



Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping

Christophe Pradal, Sarah Cohen-Boulakia, Gaëtan Heidsieck, Esther Pacitti, Francois Tardieu, Patrick Valduriez

► To cite this version:

Christophe Pradal, Sarah Cohen-Boulakia, Gaëtan Heidsieck, Esther Pacitti, Francois Tardieu, et al.. Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping. ERCIM News, 2018, Smart Farming, pp.36-37. hal-01948568

HAL Id: hal-01948568

<https://inria.hal.science/hal-01948568>

Submitted on 10 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed Management of Scientific Workflows for High-Throughput Plant Phenotyping

by Christophe Pradal (CIRAD), Sarah Cohen-Boulakia (Univ. Paris-Saclay), Gaetan Heidsieck (Inria), Esther Pacitti (Univ. Montpellier), François Tardieu (INRA) and Patrick Valduries (Inria)

High-throughput phenotyping platforms allow acquisition of quantitative data on thousands of plants required for genetic analyses in well-controlled environmental conditions. However, analysing these massive datasets and reproducing computational experiments require the use of new computational infrastructure and algorithms to scale.

Plant species will need particular characteristics to survive in the world's changing climate. To this end, plant scientists are analysing traits of interest to identify both the natural genetic variations in plants and the genetic control of their responses to environmental cues.

In the last decade, high-throughput phenotyping platforms have allowed acquisition of quantitative data on thousands of plants required for genetic analyses in well-controlled environmental conditions. These platforms in controlled con-

ditions produce huge datasets (thousands of images, environmental conditions and sensor outputs) with complex in-silico data analyses that result in the definition of new, complex variables [2]. The seven facilities of Phenome produce 200 terabytes of data annually, which are heterogeneous (images, time courses), multiscale (from the organ to the field) and originate from different sites. This infrastructure paves the way for the collection of data at an even larger scale. Indeed, farmers and breeders can

analysis of these massive datasets and the ability to reproduce large and complex in-silico experiments.

Such experiments can actually be represented and executed in an efficient and reproducible way by means of scientific workflows in which computational tasks can be chained (e.g., upload input files, preprocess the data, run various analyses and simulations, aggregate the results). OpenAlea is a scientific workflow system that provides methods and software for plant modelling at different scales [L1]. We have used it in the context of Phenome, by developing the OpenAlea Phenomenal software package dedicated to the analysis of 3-D plant architecture, whose outputs can be used for ecophysiological modelling. Phenomenal provides fully automatic workflows dedicated to the 3D reconstruction, segmentation and tracking of plant organs. It has been tested on maize, cotton, sorghum and apple trees. OpenAlea radiative models are used to estimate the light use efficiency and the in-silico crop performance in a large range of contexts.

Executing workflows on large datasets is time-consuming, and thus often incompatible with the scientist's way of working, where trial and error is an essential component. We have designed the infrastructure InfraPhenoGrid [1] to distribute the computation of workflows using the EGI/France Grilles computing facilities. EGI provides access to a grid with multiple sites, each with one or more clusters. This environment is now well suited for data-intensive computing, with different research groups collaborating at different sites. In this context, the goal is to address two critical issues in the management of plant phenotyping experiments: 1) scheduling distributed computation and

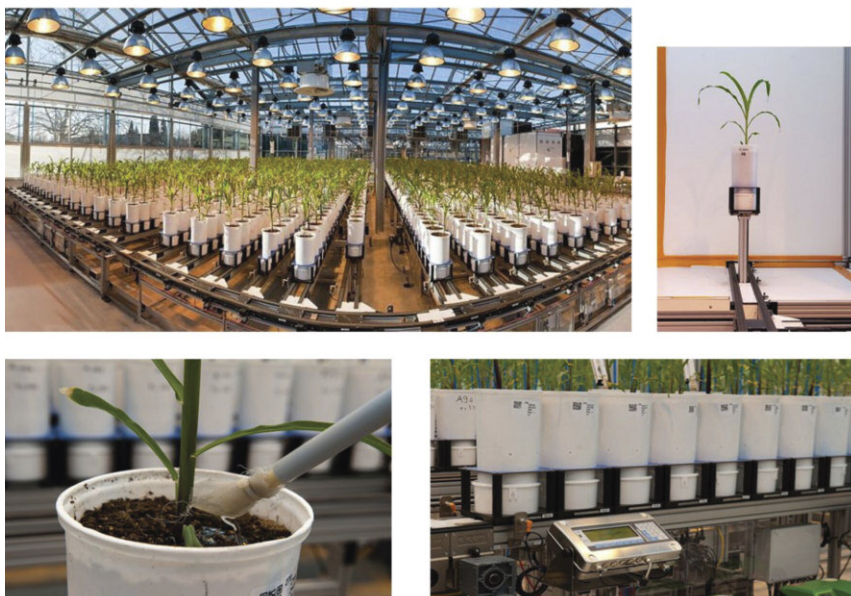


Figure 1 The Phenoarch phenotyping platform is one of the Phenome node in Montpellier. It has a capacity of 1,500 plants per experiment with a controlled environment (e.g., temperature, humidity, irrigation) and automatic imaging through time (13 images per plant per day).

ditions are essential to understand the variability of yield in the field depending on climatic scenarios [3]. Recently, national infrastructures such as Phenome have used high-throughput platforms to observe the dynamic growth of a large number of plants under in field and platform conditions. These

increasingly capture huge amounts of diverse data, which at this stage is mostly environmental, but which can involve detailed maps of yield in fields (precision agriculture), and images originating from remote sensing and drone imaging. Hence, the major problem becomes the automatic

